# Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/JP2005/017696

International filing date: 27 September 2005 (27.09.2005)

Document type: Certified copy of priority document

Document details: Country/Office: JP
Number: 2004-345392
Filing date: 30 November 2004 (30.11.2004)

Date of receipt at the International Bureau: 17 November 2005 (17.11.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)

# 日 本 国 特 許 庁
## JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

| | |
|---|---|
| 出 願 年 月 日<br>Date of Application: | ２００４年１１月３０日 |
| 出 願 番 号<br>Application Number: | 特願２００４－３４５３９２ |
| パリ条約による外国への出願に用いる優先権の主張の基礎となる出願の国コードと出願番号<br>The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is | ＪＰ２００４－３４５３９２ |
| 出 願 人<br>Applicant(s): | 松下電器産業株式会社 |

２００５年１１月 ２日

特許庁長官
Commissioner,
Japan Patent Office

中嶋 誠

【書類名】　特許願
【整理番号】　7048060194
【提出日】　平成16年11月30日
【あて先】　特許庁長官殿
【国際特許分類】　G06F 17/30
【発明者】
　【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内
　【氏名】　稲葉　光昭
【発明者】
　【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内
　【氏名】　菅野　祐司
【特許出願人】
　【識別番号】　000005821
　【氏名又は名称】　松下電器産業株式会社
【代理人】
　【識別番号】　100097445
　【弁理士】
　【氏名又は名称】　岩橋　文雄
【選任した代理人】
　【識別番号】　100103355
　【弁理士】
　【氏名又は名称】　坂口　智康
【選任した代理人】
　【識別番号】　100109667
　【弁理士】
　【氏名又は名称】　内藤　浩樹
【手数料の表示】
　【予納台帳番号】　011305
　【納付金額】　16,000円
【提出物件の目録】
　【物件名】　特許請求の範囲　1
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　9809938

【書類名】特許請求の範囲
【請求項１】
構造化文書を管理するデータベース構築装置において、
構造化文書にユニークな文書番号を割り当てるとともに構造の解析を行う入力文書解析手段と、
前記入力文書解析手段の解析結果に基づいて、構造化文書に出現する各要素名に対してユニークな要素名ＩＤを割り当てて要素名辞書に登録する要素名登録手段と、
前記入力文書解析手段の解析結果に基づいて、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名ＩＤを割り当てて祖先パス名辞書に登録する祖先パス名登録手段と、
前記入力文書解析手段の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順の情報を少なくとも含む要素出現情報を、要素名ＩＤをキーとして要素出現情報格納手段に、
着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名ＩＤをキーとして祖先パス出現情報格納手段にそれぞれ登録する出現情報登録手段と、
を有することを特徴とするデータベース構築装置。
【請求項２】
前記入力文書解析手段の解析結果に基づいて、構造化文書に出現する各属性名に対してユニークな属性名ＩＤを割り当てて属性名辞書に登録する属性名登録手段と、
着目属性の出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと分岐順の情報を少なくとも含む属性出現情報を、属性名ＩＤをキーとして格納した、属性出現情報格納手段とを有し、
前記出現情報登録手段が、前記入力文書解析手段の解析結果に基づいて、属性出現情報を作成して前記属性出現情報格納手段への登録も行うことを特徴とする請求項１に記載のデータベース構築装置。
【請求項３】
要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと属性名ＩＤと分岐順の情報を少なくとも含むテキスト出現情報を、切り出された部分文字列をキーとして格納した、テキスト出現情報格納手段とを有し、
前記出現情報登録手段が、前記入力文書解析手段の解析結果に基づいて、テキスト出現情報を作成してテキスト出現情報格納手段にも登録することを特徴とする請求項２に記載のデータベース構築装置。
【請求項４】
構造化文書を管理するデータベース検索装置において、
構造化文書に出現する各要素名に対してユニークな要素名ＩＤを登録した要素名辞書と、
前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名ＩＤを登録した祖先パス名辞書と、
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順の情報を少なくとも含む要素出現情報を、要素名ＩＤをキーとして格納した要素出現情報格納手段と、
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名ＩＤをキーとして格納した、祖先パス出現情報格納手段と、
検索式を入力するための検索条件入力手段と、
前記要素名辞書と前記祖先パス名辞書とを参照して、前記入力された検索式を内部条件式に変換する検索条件解析手段と、
前記検索条件解析手段の出力した内部条件式に従って、前記要素出現情報格納手段からの要素出現情報及び、前記祖先パス出現情報格納手段からの祖先パス出現情報から検索結果

群を求める出現情報取得手段と、
を有することを特徴とするデータベース検索装置。
【請求項５】
属性名ＩＤと対応する属性名の記録された属性名辞書と、
着目属性の出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと分岐順の情報を少なくとも含む属性出現情報を、属性名ＩＤをキーとして格納した属性出現情報格納手段とを有し、
前記検索条件解析手段が、前記要素名辞書と前記祖先パス名辞書と前記属性名辞書とを参照して、前記検索条件入力手段から入力された検索式を内部条件式に変換し、前記出現情報取得手段が、前記検索条件解析手段の出力した内部条件式に従って、前記要素出現情報格納手段からの要素出現情報、前記祖先パス出現情報格納手段からの祖先パス出現情報及び、前記属性出現情報格納手段からの属性出現情報から検索結果群を求めることを特徴とする請求項４に記載のデータベース検索装置。
【請求項６】
要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと属性名ＩＤと分岐順の情報を少なくとも含むテキスト出現情報を、切り出された部分文字列をキーとして格納した、テキスト出現情報格納手段とを有し、
前記出現情報取得手段が、前記検索条件解析手段の出力した内部条件式に従って、前記要素出現情報格納手段からの要素出現情報、前記祖先パス出現情報格納手段からの祖先パス出現情報、前記属性出現情報格納手段からの属性出現情報及び、前記テキスト出現情報格納手段からのテキスト出現情報から検索結果群を求めることを特徴とする請求項５に記載のデータベース検索装置。
【請求項７】
前記出現情報取得手段は、前記要素出現情報格納手段における指定要素名ＩＤのエントリ数と、前記祖先パス出現情報格納手段における指定祖先パス名ＩＤのエントリ数の大小を比較し、少ない方の出現情報を参照するようにして検索結果群を求めることを特徴とする請求項４乃至６のいずれかに記載のデータベース検索装置。
【請求項８】
構造化文書を管理するデータベース構築方法において、
構造化文書にユニークな文書番号を割り当てるとともに構造の解析を行うステップと、
前記解析結果に基づいて、構造化文書に出現する各要素名に対してユニークな要素名ＩＤを割り当てて要素名辞書に登録するステップと、
前記解析結果に基づいて、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名ＩＤを割り当てて祖先パス名辞書に登録するステップと、
前記解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順の情報を少なくとも含む要素出現情報を、要素名ＩＤをキーとして要素出現情報格納部に、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名ＩＤをキーとして祖先パス出現情報格納部にそれぞれ登録するステップと、を有することを特徴とするデータベース構築方法。
【請求項９】
構造化文書を管理するデータベース検索方法において、
構造化文書に出現する各要素名に対してユニークな要素名ＩＤを登録した要素名辞書と、
前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名ＩＤを登録した祖先パス名辞書と、
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順の情報を少なくとも含む要素出現情報を、要素名ＩＤをキーとして格納した要素出現情報格納部と、
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名ＩＤをキーとして

格納した、祖先パス出現情報格納部と、
検索式を入力するためのステップと、
前記要素名辞書と前記祖先パス名辞書とを参照して、前記入力された検索式を内部条件式に変換するステップと、
前記内部条件式に従って、前記要素出現情報格納部からの要素出現情報及び、前記祖先パス出現情報格納部からの祖先パス出現情報から検索結果群を求めるステップと、
を有することを特徴とするデータベース検索方法。

【書類名】明細書
【発明の名称】データベース構築装置及びデータベース検索装置
【技術分野】
【０００１】
本発明は、ＸＭＬなどの構造化文書を、管理するデータベース構築検索技術に関するもので、特に、大量の構造化文書群の中から指定した論理構造を持つ文書を効率良く検索するデータベース構築装置及びデータベース検索装置に関するものである。
【背景技術】
【０００２】
従来の構造化文書を管理する装置としては、例えば特許文献１が知られている。
【０００３】
以下、従来例の概要について図を参照しながら説明する。図２４は従来の構造化文書管理装置の構成図である。登録対象の構造化文書は構造化文書入力手段２４０２から入力し、構造解析手段２４０７によって解析され、木構造を得る。構造情報作成手段２４０８によって、各要素のタグ名（要素名）には名称ＩＤが割り振られて名称ＩＤテーブル格納手段２４１８に格納される。また、各要素のパス名称（最上位階層から順にタグ名を連ねて記述した文字列）にはパス名称ＩＤが割り振られて、パス名称インデックス格納手段２４１６に、各要素のパス階層（パス名称の各階層の出現順序（同じ親要素を持つ同じタグ名の要素の中で何番目に出現した要素か）を連ねて記述した文字列）にはパス階層ＩＤが割り当てられて、パス階層インデックス格納手段２４１７に格納に、実体（テキスト）を持つ要素（要素実体）の場合は、各要素実体に対し、検索単位を一意に表す符合（検索単位識別子と呼ぶ）が割り当てられ、この検索単位識別子をキーとして、文書番号，パス名称ＩＤ，パス階層ＩＤ，名称ＩＤの組が要素管理テーブル格納手段２４１５にそれぞれ格納される。図２５は要素管理テーブルの例を示したものである。
【０００４】
次に、文字列索引作成手段２４０９は、各要素実体の内容の文字列に対して、予め定めた文字数の文字連鎖を取り出す。この文字連鎖について、該当する検索単位識別子、および該文字連鎖先頭文字がその要素内容において何番目の文字かを表す番号（文字位置番号）を文字列索引格納手段２４１９に登録する。図２６は文字列索引の例の一部を示した図である。図２６の２６０１は「検索単位識別子が“１”の要素の文字列中に“構造”という文字連鎖が先頭から“１”文字目の位置から存在する」ということを表している。
【０００５】
次に、このようにして格納されたデータを用いた検索の概要を説明する。図２７は検索条件として「パス名称が“／論文／書誌／タイトル”である要素に“構造化”という文字列が含まれる文書」が与えられた場合の処理を図に示したものである。検索条件解析手段２４１０は、パス名称インデックス２４１６を参照し、検索条件のパス名称をパス名称ＩＤ“Ｎ２”に変換する。次に文字列索引検索手段２４１１は“構造化”から2文字連鎖“構造”と“造化”を取り出す。文字列索引を参照し、“構造”と“造化”が連続して出現し、かつ検索単位識別子が同一なものを求め、その検索単位識別子を抽出する。図２７では検索単位識別子“１”と“８”が文字列索引検索結果群として返っている。次に、構造照合手段２４１２が検索条件の構造指定を満たす最終的な検索結果を求める。文字列索引検索結果群として得られた検索単位識別子をキーにして、要素管理テーブルを参照し、パス名称ＩＤが“Ｎ２”に一致するものだけを最終的な検索結果とする。
【０００６】
その他、タグ名を指定した検索条件であれば、要素管理テーブルの名称ＩＤが指定タグ名の名称ＩＤと一致するものだけを最終的な検索結果とし、パス名称とパス階層を共に指定した検索条件であれば、要素管理テーブルのパス名称ＩＤが指定したパス名称のパス名称ＩＤと一致し、かつパス階層ＩＤが指定したパス階層のパス階層ＩＤと一致するものだけを最終的な検索結果とする。
【特許文献１】特開２００２－２０２９７３号公報

【発明の開示】
【発明が解決しようとする課題】
【０００７】
しかしながら、前記従来の構成では、まず文字列索引を参照して指定された文字列の出現する検索単位識別子を求めた後、検索単位識別子が指定された構造条件を満たすかどうかを、要素管理テーブルを参照して判定する。そのため、文字列検索条件の指定は必須であり、構造条件だけを指定した検索を行うことができない（行うためには全ての検索単位識別子について構造条件を満たすかどうかを判定しなければならず、要素管理テーブル全体をサーチしなくてはならないため、効率が非常に悪い）という課題を有していた。また、文字列索引は要素実体の内容文字列に対してのみ作成されるため、属性値に対しては文字列検索を行うことができないという課題を有していた。
【０００８】
本発明は、前記従来の課題を解決するもので、文字列検索条件と構造条件をともに指定した場合だけでなく、文字列検索条件を伴わない構造だけを指定した様々な検索条件に対しても、所望の文書を効率良く検索することが可能なデータベース構築検索装置を提供することを目的とする。さらに、要素内のテキスト文字列だけでなく、属性値に対しても文字列検索が可能なデータベース構築検索装置を提供することを目的とする。
【課題を解決するための手段】
【０００９】
前記従来の課題を解決するために、本発明のデータベース構築検索装置は、要素の出現する文書番号，文字位置，文字数，祖先パス名ＩＤ，分岐順の情報を、要素名ＩＤをキーにして格納した要素出現情報格納手段と、要素の出現する文書番号，文字位置，文字数，要素名ＩＤ，分岐順の情報を、その要素の祖先パス名ＩＤをキーにして格納した祖先パス出現情報格納手段と、属性の出現する文書番号，文字位置，文字数，要素名ＩＤ，祖先パス名ＩＤ，分岐順の情報を、属性名ＩＤをキーにして格納した属性出現情報格納手段と、を備え、文字列検索条件を伴わない構造条件のみが指定された様々な検索式に対して、検索式の形と出現情報のエントリ数を考慮して、適切な出現情報を選択して使用することにより所望の文書を効率よく求めることができる。
【００１０】
また、要素実体のテキストから切り出した部分文字列、および要素のもつ属性の値から切り出した部分文字列に関して、出現する文書番号，文字位置，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順の情報を、部分文字列をキーにして格納したテキスト出現情報格納手段を備え、要素内のテキストに対する文字列検索だけでなく、属性値に対しても文字列検索が可能となる。
【発明の効果】
【００１１】
本発明のデータベース構築検索装置によれば、文字列検索条件と構造条件をともに指定した検索条件のみならず、構造だけを指定した様々な検索条件に対しても、所望の論理構造を持つ文書を効率よく検索することが可能となる。また、要素実体のテキスト文字列に対してだけでなく、属性値に対しても文字列検索を行うことが可能となる。
【発明を実施するための最良の形態】
【００１２】
以下本発明の実施の形態について、図面を参照しながら説明する。
【００１３】
（実施の形態１）
図１は、本発明の実施の形態１におけるデータベース構築検索装置の構成図である。
【００１４】
図１において、１０１はデータベースに登録する構造化文書群、１０２は入力された構造化文書群１０１の各文書についてユニークな文書番号を割り振るとともに論理構造の解析を行う入力文書解析手段、１０３は入力文書解析手段１０２の解析結果から、文書に出

現する要素名に対してユニークな識別子（以下、要素名ＩＤと呼ぶ）を割り当てて要素名辞書１０７に登録する要素名登録手段、１０４は入力文書解析手段１０２の解析結果から、文書に出現する祖先パス名（着目要素の祖先要素の要素名を最上位階層から順にスラッシュで区切って並べた文字列で、着目要素自身の要素名は含まない）に対してユニークな識別子（以下、祖先パス名ＩＤと呼ぶ）を割り当てて祖先パス名辞書１０８に登録する祖先パス名登録手段、１０５は入力文書解析手段１０２の解析結果から、文書に出現する属性名に対してユニークな識別子（以下、属性名ＩＤと呼ぶ）を割り当てて属性名辞書１０９に登録する属性名登録手段、１０６は入力文書解析手段１０２の解析結果から、出現位置索引１１０の要素出現情報格納手段１１１、祖先パス出現情報格納手段１１２、属性出現情報格納手段１１３、テキスト出現情報格納手段１１４に四種の出現情報を登録する出現情報登録手段、１０７は要素名ＩＤとそれに対応する要素名が記録された要素名辞書、１０８は祖先パス名ＩＤとそれに対応する祖先パス名が記録された祖先パス名辞書、１０９は属性名ＩＤとそれに対応する属性名が記録された属性名辞書、１１０は要素出現情報格納手段１１１、祖先パス出現情報格納手段１１２、属性出現情報格納手段１１３、テキスト出現情報格納手段１１４、の四種の出現情報が格納されている出現位置索引格納手段、１１１は各要素の出現する文書番号，文字位置，文字数，祖先パス名ＩＤ，分岐順の情報を、要素名ＩＤをキーにして格納した要素出現情報格納手段、１１２は各要素の出現する文書番号，文字位置，文字数，要素名ＩＤ，分岐順の情報を、その要素の祖先パス名ＩＤをキーにして格納した、祖先パス出現情報格納手段、１１３は各属性の出現する文書番号，文字位置，文字数，要素名ＩＤ，祖先パス名ＩＤ，分岐順の情報を、属性名ＩＤをキーにして格納した属性出現情報格納手段、１１４は要素内のテキストから切り出した部分文字列、および要素のもつ属性の値から切り出した部分文字列に関して、出現する文書番号，文字位置，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順の情報を、部分文字列をキーにして格納したテキスト出現情報格納手段、１１６は検索式１１５を受け付ける検索条件入力手段、１１７は、検索条件入力手段１１６に与えられた検索式を解析し、内部条件に変換して出現情報取得手段１１８に出力する検索条件解析手段、１１８は検索条件解析手段１１７の出力した内部条件にしたがって、出現位置索引１１０に格納された四種の出現情報から適切な情報を選択して取得し、検索条件にマッチする結果データ集合を求める出現情報取得手段、１１９は結果データ集合を適切な形式で検索結果１２０として出力する検索結果出力手段である。

【００１５】

上記のように構成されたデータベース構築検索装置の動作について説明する。

【００１６】

はじめに、文書登録（データベース構築）処理に関して具体例を挙げて説明する。図２２は、文書の登録処理の流れを表す図である。

【００１７】

まず、ステップ２２０１において、入力文書解析手段１０２は、構造化文書群１０１から構造化文書を１つ読み込んで、ユニークな文書番号を割り振る。

【００１８】

次に、ステップ２２０２で、この文書の論理構造を解析する。図２は構造化文書の一例である（構造化文書群１０１には、このような文書が複数含まれる）。図２に示した構造化文書は、最上位階層にbook要素を持ち、book要素はtitle要素と２つのchapter要素を含んでいる。title要素は要素実体の文字列“文書検索”を含み、１つ目のchapter要素は別のtitle要素と２つのsection要素および属性値が“歴史”であるkeyword属性を持つ構造を持っている。図２に示す構造化文書を入力文書解析手段１０２によって解析した結果得られる木構造は、図３のようになる。図３において、四角い枠は要素３０１～３０３を表し、枠内に記された文字列は要素名３０４を示している。また、楕円の点線枠は属性３０５を表し、枠内に記された文字列は属性名３０６を示している。木構造の最上位階層の要素３０１から着目要素に至る経路の途中に存在する要素（祖先要素）の要素名をスラッシュで区切って順に並べたものはパス名と呼ばれる。パス名のうちの末尾部分（＝着目要素

自身の要素名）を除いた部分を「祖先パス名」と呼ぶことにする。図７は、図３の網掛けを施した要素３０２に関するパス名７０１，祖先パス名７０２，要素名７０３を示したものである。また、図３において、要素の右肩に記された“１／２／３”などの文字列は、パス名中の各要素について、同じ親要素を持つ同じ要素名の要素の中で何番目に出現したかの順を示す番号を並べたもので、これを「分岐順」３０７と呼ぶ。図３の網掛けを施した要素３０２とその左隣の要素３０３とは、パス名は同じであるが分岐順３０７，３０８は異なっている。

【００１９】

次に、入力文書解析手段１０２の解析結果をうけて、当該文書に出現する各要素について以下の処理を繰り返す。

【００２０】

ステップ２２０３において、要素名登録手段１０３は、着目要素の要素名が要素名辞書１０７に登録済みかどうかを調べ、登録済みであれば対応する要素名ＩＤを取得し、登録されていなければ新たに要素名ＩＤ（＞０）を割り当てて要素名辞書１０７に登録する。

【００２１】

ステップ２２０４で、祖先パス名登録手段１０４は、着目要素の祖先パス名が祖先パス名辞書１０８に登録済みかどうかを調べ、登録済みであれば対応する祖先パス名ＩＤを取得し、登録されていなければ新たに祖先パス名ＩＤ（＞０）を割り当てて祖先パス名辞書１０８に登録する。

【００２２】

もし、着目要素が属性を持っているならば、ステップ２２０５～ステップ２２０６において、属性名登録手段１０５は、着目要素の各属性の属性名が属性名辞書１０９に登録済みかどうかを調べ、登録済みであれば対応する属性名ＩＤを取得し、登録されていなければ新たに属性名ＩＤ（＞０）を割り当てて属性名辞書１０９に登録する。図４、図５、図６はそれぞれ構造化文書（図２）の登録処理が終わった後の要素名辞書１０７、祖先パス名辞書１０８、属性名辞書１０９の内容の例を示している。

【００２３】

ステップ２２０７において、出現情報登録手段１０６は、着目要素に関する要素出現情報を、要素名ＩＤをキーとして要素出現情報格納手段１１１に登録する。要素出現情報は、文書番号，着目要素（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字位置および文字数，祖先パス名ＩＤ、分岐順の５種類の値の組から構成される。なお、「文字位置」は、図８に示すように、タグを除く当該文書内の全てのテキストをつなげた文字列において先頭から何文字目に当たるかで表す。その一例を図９に示すもので、図３の網掛けを施した要素３０２に関する要素出現情報の内容で、要素名ＩＤが４（＝要素名がsection）である要素が文書番号１の文書の１１５文字目から始まる長さ４０文字の要素実体を含んでいて、その祖先パス名ＩＤが３（＝祖先パス名が/book/chapter）で分岐順が1/2/3であることを表している。

【００２４】

ステップ２２０８において、出現情報登録手段１０６は、着目要素に関する祖先パス出現情報（すなわち、文書番号，着目要素（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字位置および文字数，要素名ＩＤ，分岐順の５種類の値の組）を、祖先パス名ＩＤをキーとして祖先パス出現情報格納手段１１２に登録する。図１０は、図３の網掛けを施した要素３０２に関する祖先パス出現情報の内容を示している。図９と図１０を比較してわかるように、同一要素に関する要素出現情報と祖先パス出現情報は、キーとなる項目が要素名ＩＤであるか祖先パス名ＩＤであるかという点が異なるだけである。

【００２５】

もし、着目要素が属性を持っているならば、ステップ２２０９～ステップ２２１０において、出現情報登録手段１０６は着目要素の各属性に関する属性出現情報を、属性名ＩＤをキーとして属性出現情報格納手段１１３に登録する。属性出現情報は、文書番号、属性値の先頭文字位置および文字数、祖先パス名ＩＤ、要素名ＩＤ、分岐順の６種類の値の組

から構成される。図１１は、図３の網掛けを施した要素３０２の「update」属性３０５に関する属性出現情報の内容である。その内容は、属性名ＩＤが２（＝属性名がupdate）の属性が文書番号１の文書の１１５文字目から始まる長さ６文字の属性値を持ち、属性の所属する要素の祖先パス名ＩＤが３（＝祖先パス名が/book/section）、要素名ＩＤが４（＝要素名がsection）、分岐順が1/2/3であることを示している。なお、属性出現情報において、属性値の先頭文字位置は、図１１に示すように、仮想的に着目要素（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字位置と同じであるとする。

【００２６】

ステップ２２１１において、出現情報登録手段１０６は、着目要素の実体内容のテキストから部分文字列の切り出しを行い、テキスト出現情報を、切り出された部分文字列をキーとしてテキスト出現情報格納手段１１４に登録する。ただし、属性値ではないので、属性名ＩＤには常に０を格納する。テキスト出現情報は、文書番号、切り出された部分文字列の先頭文字位置、祖先パス名ＩＤ、要素名ＩＤ、属性名ＩＤ、分岐順の６種類の値の組から構成される。

【００２７】

もし、着目要素が属性を持っているならば、ステップ２２１２～ステップ２２１３において、出現情報登録手段１０６は、着目要素が持つ各属性の属性値文字列から部分文字列の切り出しを行い、テキスト出現情報格納手段１１４に部分文字列をキーとして登録する。なお、属性出現情報と同様に、属性値は図１０に示すような位置に仮想的に出現しているとして、文字位置を算出する。また、ステップ２２１２ではステップ２２１１の場合とは異なり、属性名ＩＤには着目している属性の属性名ＩＤ（＞０）を格納する。図１２は図３の網掛けを施した要素３０２のテキストおよび「update」属性３０５の属性値についてのテキスト出現情報の一部である。図１２において、１２０１は、“極大”という部分文字列が文書番号１の文書の１１８文字目に現れ、祖先パス名ＩＤが３（＝祖先パス名が/book/section）、要素名ＩＤが４（要素名がchapter）、分岐順が1/2/3であるような要素の要素実体に含まれている（属性名ＩＤが０であることからわかる）ことを表している。また１２０２は、“００”という部分文字列が文書番号１の文書の１１８文字目に現れ、祖先パス名ＩＤが３（＝祖先パス名が/book/section）、要素名ＩＤが４（＝要素名がchapter）、分岐順が1/2/3であるような要素に属する属性名ＩＤが２（＝属性名がupdate）の属性の属性値に含まれていることを表している。

【００２８】

ステップ２２１４で、この文書に出現する全ての要素について処理が終わったかどうかを調べ、もし未処理の要素が残っていればステップ２２０３に戻って処理を繰り返す。

【００２９】

ステップ２２１５で、全ての入力文書に対して処理が終わったかどうかを調べ、未処理の文書が残っていればステップ２２０１に戻って処理を繰り返す。

【００３０】

以上のようにして、文書登録（データベース構築）処理が完了する。

【００３１】

続いて、登録済みの文書群に対する検索処理に関して説明する。

【００３２】

図２１は、検索条件入力手段１１６に与えられる検索式１１５の例をいくつか示したもので、これらの式はＷ３Ｃ（World Wide Web Consortium）の勧告として公開されているＸＰａｔｈ言語（詳細な仕様はhttp://www.w3.org/TR/xpath に記載されている）で記述されている。

【００３３】

図２１のそれぞれのＸＰａｔｈ式は、次のような意味を表している。検索式２１０１は「最上位階層のbook要素の子のchapter要素の子であるtitle要素」を、検索式２１０２は「最上位階層のbook要素の子のchapter要素のいずれかの子要素」を、検索式２１０３は「いずれかの階層にあるtitle要素」を、検索式２１０４は「最上位階層のbook要素の子

のchapter要素の子の２番目のsection要素」を、検索式２１０５は「最上位階層のbook要素の子のchapter要素の子のsection要素のupdate属性」を、検索式２１０６は「最上位階層のbook要素の子のchapter要素の子のsection要素で、かつ要素実体内容に“極大単語”という文字列を含む要素」を、検索式２１０７は「最上位階層のbook要素の子のchapter要素の子のsection要素のupdate属性で、かつその属性値に“２００４”という文字列を含む」をそれぞれ表している。

【００３４】

次に、それぞれの検索式に対して、どのような検索処理が行われるのかを順に説明する。

【００３５】

（検索式２１０１の場合）

図２３は、検索式２１０１の場合の検索処理の流れを示している。

【００３６】

ステップ２３０１において、検索条件入力手段１１６に入力された検索式２１０１は、検索条件解析手段１１７で解析される。

【００３７】

ステップ２３０２において、検索条件解析手段１１７は、検索式２１０１を解析し、要素名辞書１０７、祖先パス名辞書１０８を参照して内部条件「祖先パス名ＩＤ＝３かつ要素名ＩＤ＝２」に変換し、出現情報取得手段１１８に出力する。

【００３８】

次に、ステップ２３０３からステップ２３０５において、出現情報取得手段１１８は、出現位置索引格納手段１１０を参照し、要素出現情報格納手段１１１における要素名ＩＤ＝２のエントリ数Ｎと祖先パス出現情報格納手段１１２における祖先パス名ＩＤ＝３のエントリ数Ｍとを比較し、少ない方を選択する。図１３は、要素出現情報格納手段１１１における要素名ＩＤ＝２のエントリ１３０１、図１４は祖先パス出現情報格納手段１１２における祖先パス名ＩＤ＝３のエントリ１４０１の例で、この場合はＮ＝８，Ｍ＝１２であるから図１３の要素出現情報格納手段１１１を選ぶことになる。

【００３９】

そして、ステップ２３０６において、要素出現情報格納手段１１１の要素名ＩＤ＝２のエントリ１３０１から１つ取得し、ステップ２３０７で、このエントリの祖先パス名ＩＤが３であるかどうかを調べ、もし祖先パス名ＩＤが３であればステップ２３０８でこのエントリのデータを結果データ集合１３０２に追加する。結果データ集合の各データは例えば（文書番号，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順）のような形式である。

【００４０】

ステップ２３０９でＮエントリ全てについて処理したか調べ、まだ未処理のエントリがあればステップ２３０６に戻って処理を繰り返す。

【００４１】

ステップ２３０５において、もしＭ≦Ｎであれば、図１４のように祖先パス出現情報格納手段１１２における祖先パス名ＩＤ＝３の各エントリ１４０１を調べ、要素名ＩＤが２であるものを求め（ステップ２３１０～ステップ２３１３）結果データ集合１４０２に追加する。

【００４２】

ステップ２３１４で、求められた結果データ集合を検索結果出力手段１１９に出力する。最後に検索結果出力手段１１９は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【００４３】

このように、検索式２１０１に対しては、要素出現情報格納手段１１１における指定した要素名ＩＤのエントリから指定した祖先パス名ＩＤを持つものを選ぶという処理と、祖先パス出現情報格納手段１１２における指定した祖先パス名ＩＤのエントリから指定した要素名ＩＤを持つものを選ぶという2種類の処理のどちらか、エントリ数の少ないほうを

選ぶことによって、検索対象構造化文書群の論理構造の特性に応じて処理量を抑えることができるため、所望の文書を効率よく検索することができる。

【００４４】

（検索式２１０２の場合）

検索条件入力手段１１６に入力された検索式２１０２は、検索条件解析手段１１７で解析される。検索条件解析手段１１７は、検索式２１０２を解析し、祖先パス名辞書１０８を参照して内部条件「祖先パス名ＩＤ＝３」に変換し、出現情報取得手段１１８に出力する。出現情報取得手段１１８は、出現位置索引１１０を参照し、図１５のように祖先パス出現情報格納手段１１２における祖先パス名ＩＤ＝３の全てのエントリ１５０１を求め、例えば（文書番号，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順）のような形式で結果データ集合１５０２として検索結果出力手段１１９に出力する。検索結果出力手段１１９は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【００４５】

このように、検索式２１０２に対しては、祖先パス出現情報格納手段１１２における指定した祖先パス名ＩＤのエントリを取得するだけで良いため、所望の文書を効率よく検索することができる。

【００４６】

（検索式２１０３の場合）

検索条件入力手段１１６に入力された検索式２１０３は、検索条件解析手段１１７で解析される。検索条件解析手段１１７は、検索式２１０３を解析し、要素名辞書１０７を参照して内部条件「要素名ＩＤ＝２」に変換し、出現情報取得手段１１８に出力する。出現情報取得手段１１８は、出現位置索引格納手段１１０を参照し、図１６のように要素出現情報格納手段１１１における要素名ＩＤ＝２の全てのエントリ１６０１を求め、例えば（文書番号，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順）のような形式で結果データ集合１６０２を検索結果出力手段１１９に出力する。検索結果出力手段１１９は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【００４７】

このように、検索式２１０３に対しては、要素出現情報格納手段１１１における指定した要素名ＩＤのエントリを取得するだけで良いため、所望の文書を効率よく検索することができる。

【００４８】

（検索式２１０４の場合）

検索条件入力手段１１６に入力された検索式２１０４は、検索条件解析手段１１７で解析される。検索条件解析手段１１７は、検索式２１０４を解析し、要素名辞書１０７、祖先パス名辞書１０８を参照して内部条件「祖先パス名ＩＤ＝３かつ要素名ＩＤ＝４かつ分岐順＝”＊／＊／２”」に変換し、出現情報取得手段１１８に出力する。分岐順のアスタリスク「＊」の部分はどんな数字でもマッチすることを表す。出現情報取得手段１１８は、出現位置索引１１０を参照し、要素出現情報格納手段１１１における要素名ＩＤ＝４のエントリ数Ｎと祖先パス出現情報格納手段１１２における祖先パス名ＩＤ＝３のエントリ数Ｍとを比較し、少ないほうを選択する。

【００４９】

もし、Ｍ≦Ｎであれば、図１７に示すように祖先パス出現情報格納手段１１２における祖先パス名ＩＤ＝３の各エントリ１７０１を調べ、要素名ＩＤが４であり、かつ分岐順が”＊／＊／２”であるエントリのデータを結果データ集合１７０２として、例えば（文書番号，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順）のような形式で検索結果出力手段１１９に出力する。もし、Ｍ＞Ｎならばば要素パス出現情報格納手段１１１における要素名ＩＤ＝４の各エントリを調べ、祖先パス名ＩＤが３であり、かつ分岐順が“＊／＊／２”であるエントリのデータを結果データ集合１７０２として検索結果出力手段１１９に出力する。

【００５０】
最後に検索結果出力手段１１９は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。
【００５１】
このように、検索式２１０４に対しては、要素出現情報格納手段１１１における指定した要素名ＩＤのエントリから指定した祖先パス名ＩＤと分岐順を持つものを選ぶという処理と、祖先パス出現情報格納手段１１２における指定した祖先パス名ＩＤのエントリから指定した要素名ＩＤと分岐順を持つものを選ぶという2種類の処理のどちらか、エントリ数の少ないほうを選ぶことによって、処理量を減らすことが可能となり、所望の文書を効率よく検索することができる。
【００５２】
（検索式２１０５の場合）
検索条件入力手段１１６に入力された検索式２１０５は、検索条件解析手段１１７で解析される。検索条件解析手段１１７は、検索式２１０５を解析し、要素名辞書１０７、祖先パス名辞書１０８、属性名辞書１０９を参照して内部条件「祖先パス名ＩＤ＝３かつ要素名ＩＤ＝４かつ属性名ＩＤ＝２」に変換し、出現情報取得手段１１８に出力する。出現情報取得手段１１８は、出現位置索引１１０を参照し、図１８のように属性出現情報格納手段１１３における属性名ＩＤ＝２の各エントリ１８０１を調べ、祖先パス名ＩＤが３であり、要素名ＩＤが４であればそのエントリのデータを例えば（文書番号，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順）のような形式で結果データ集合１８０２として検索結果出力手段１１９に出力する。最後に、検索結果出力手段１１９は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。
【００５３】
このように、検索式２１０５に対しては、属性出現情報格納手段１１３における指定した属性名ＩＤのエントリから指定した祖先パス名ＩＤと要素名ＩＤを持つものを選ぶことによって、所望の文書を検索することが可能となる。
【００５４】
（検索式２１０６の場合）
検索条件入力手段１１６に入力された検索式２１０６は、検索条件解析手段１１７で解析される。検索条件解析手段１１７は、検索式２１０６を解析し、要素名辞書１０７、祖先パス名辞書１０８を参照して内部条件「祖先パス名ＩＤ＝３かつ要素名ＩＤ＝４かつ要素内に文字列“極大単語”を含む」に変換し、出現情報取得手段１１８に出力する。出現情報取得手段１１８は、出現位置索引格納手段１１０を参照し、図１９のようにテキスト出現情報格納手段１１４における“極大”のエントリ１９０１と“単語”のエントリ１９０２の間の連接演算を行う。その際、文書番号が同一であることと“単語”が“極大”の２文字後方に位置することだけでなく、祖先パス名ＩＤが３、かつ要素名ＩＤが４、かつ属性名ＩＤが０、かつ分岐順が同一であるというチェックも行い条件を満たすものを、例えば（文書番号，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順）のような形式で結果データ集合１９０３として検索結果出力手段１１９に出力する。検索結果出力手段１１９は、求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。
【００５５】
このように、検索式２１０６に対しては、テキスト出現情報格納手段１１４における部分文字列のエントリ同士の連接演算の際に、祖先パス名ＩＤおよび要素名ＩＤが指定した値であって、分岐順が同一であり，かつ属性名ＩＤが０であるものを選ぶことによって、所望の文書を検索することが可能となる。
【００５６】
（検索式２１０７の場合）
検索条件入力手段１１６に入力された検索式２１０７は、検索条件解析手段１１７で解析される。検索条件解析手段１１７は、検索式２１０７を解析し、要素名辞書１０７、祖

先パス名辞書１０８、属性名辞書１０９を参照して内部条件「祖先パス名ＩＤ＝３かつ要素名ＩＤ＝４かつ属性名ＩＤ＝２かつ属性値に文字列“２００４”を含む」に変換し、出現情報取得手段１１８に出力する。出現情報取得手段１１８は、出現情報取得手段１１８は、出現位置索引１１０を参照し、図２０のようにテキスト出現情報格納手段１１４における“２０”のエントリ２００１と“０４”のエントリ２００２の間の連接演算を行う。その際、文書番号が同一であることと“２０”が“０４”の２文字後方に位置することだけでなく、祖先パス名ＩＤが３、かつ要素名ＩＤが４、かつ属性名ＩＤが２、かつ分岐順が同一であるというチェックも行い、条件を満たすものを、例えば（文書番号，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順）のような形式で結果データ集合２００３として検索結果出力手段１１９に出力する。検索結果出力手段１１９は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【００５７】

このように、検索式２１０７に対しては、テキスト出現情報格納手段１１４における部分文字列のエントリ同士の連接演算の際に、祖先パス名ＩＤおよび要素名ＩＤが指定した値であって、分岐順が同一であり，かつ属性名ＩＤが指定した値（＞０）であるものを選ぶことによって、所望の文書を検索することが可能となる。

【００５８】

以上説明したように、要素の出現情報を、要素名ＩＤをキーにして格納した要素出現情報格納手段と、要素の出現情報をその要素の祖先パス名ＩＤをキーにして格納した祖先パス名出現情報格納手段と、属性の出現情報を、属性名ＩＤをキーにして格納した属性出現情報格納手段とを設けることにより、構造条件だけを指定した検索式に対しても効率よく所望の文書を検索することができる。また、要素実体のテキスト文字列および要素の持つ属性の属性値から切り出された部分文字列の出現情報を格納したテキスト出現情報格納手段を設けることにより、要素実体のテキストに対してだけでなく属性値に対しても文字列検索を行うことができる。

【００５９】

なお、データベース構築処理において、要素実体や属性値から固定長の２文字連鎖で部分文字列の切り出しを行うと説明したが、他の切り出し方法、例えば特開平８－２４９３５４号公報「文書検索装置および単語索引作成方法および文書検索方法」に記載の方法等でも構わない。

【００６０】

また、データベース検索処理において、検索条件式をＸＰａｔｈ式で与えるとして説明したが、同様の意味を持つ他のクエリ言語であっても本発明を適用することは可能である。

【産業上の利用可能性】

【００６１】

本発明のデータベース構築検索装置は、技術文献、特許文献、新聞、雑誌等、あらゆる電子化された構造化文書に対して広範に利用することが可能である。

【図面の簡単な説明】

【００６２】

【図１】本発明の実施の形態１の構成を示すブロック図

【図２】本発明の実施の形態１における、登録検索対象となる構造化文書の一例を示す図

【図３】本発明の実施の形態１における、構造化文書の論理構造を解析した結果である木構造の一例を示す図

【図４】発明の実施の形態１における、要素名辞書の内容の一例を示す図

【図５】本発明の実施の形態１における、祖先パス名辞書の内容の一例を示す図

【図６】本発明の実施の形態１における、属性名辞書の内容の一例を示す図

【図７】本発明の実施の形態１における、祖先パス名の説明に用いる図

【図８】本発明の実施の形態１における、文字位置の説明に用いる図

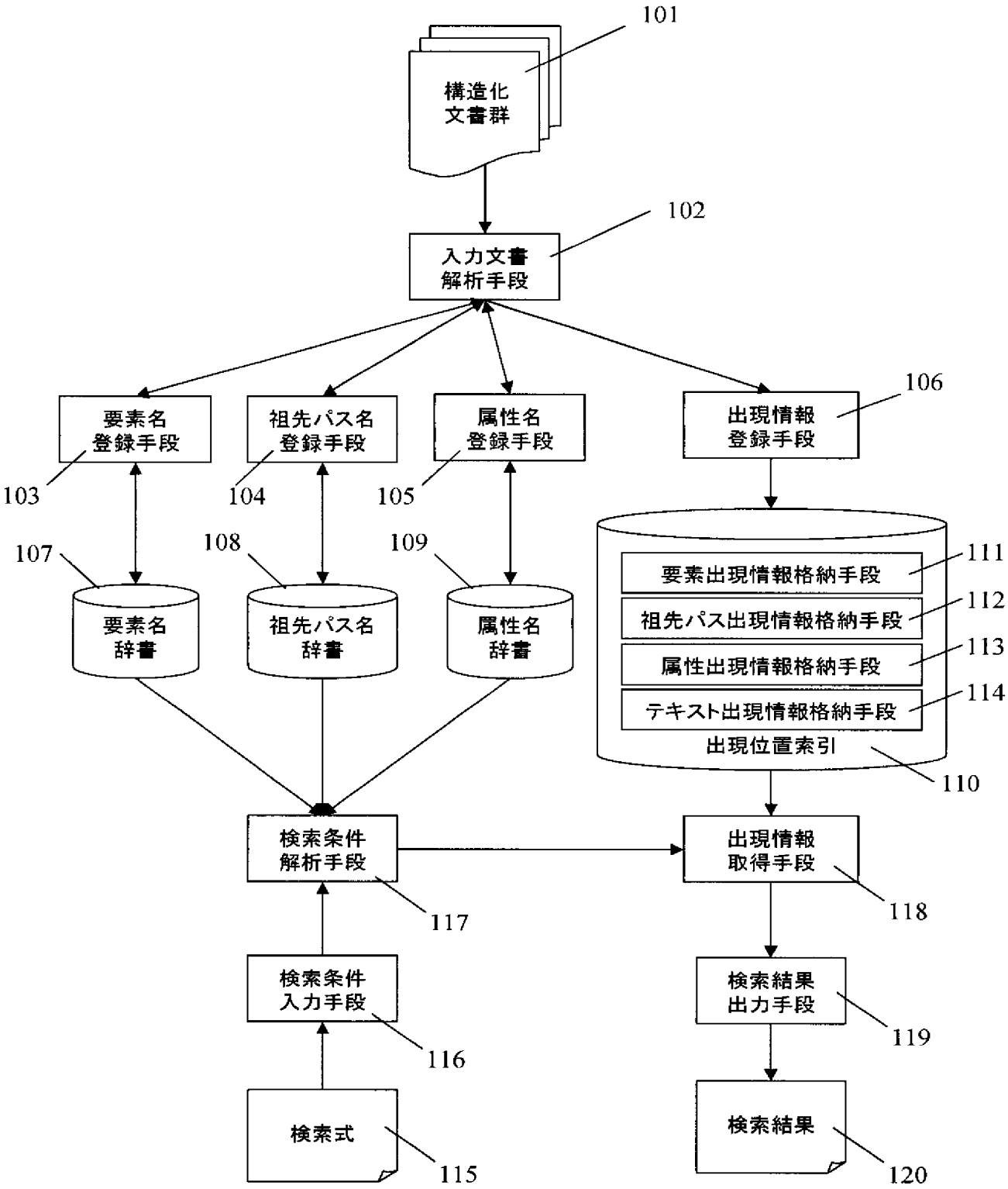
【図９】本発明の実施の形態１における、要素出現情報の説明に用いる図
【図１０】本発明の実施の形態１における、祖先パス出現情報の説明に用いる図
【図１１】本発明の実施の形態１における、属性出現情報の説明に用いる図
【図１２】本発明の実施の形態１における、テキスト出現情報の説明に用いる図
【図１３】本発明の実施の形態１における、検索処理の説明に用いる図
【図１４】本発明の実施の形態１における、検索処理の説明に用いる図
【図１５】本発明の実施の形態１における、検索処理の説明に用いる図
【図１６】本発明の実施の形態１における、検索処理の説明に用いる図
【図１７】本発明の実施の形態１における、検索処理の説明に用いる図
【図１８】本発明の実施の形態１における、検索処理の説明に用いる図
【図１９】本発明の実施の形態１における、検索処理の説明に用いる図
【図２０】本発明の実施の形態１における、検索処理の説明に用いる図
【図２１】本発明の実施の形態１における、検索式の例を示す図
【図２２】本発明の実施の形態１における、文書登録処理の手順を示す流れ図
【図２３】本発明の実施の形態１における、検索処理の手順を示す流れ図
【図２４】従来の技術における構造化文書管理装置の構成図
【図２５】従来の技術における要素管理テーブルの例を示す図
【図２６】従来の技術における文字列索引の例の一部を示す図
【図２７】従来の技術における検索処理の図

【符号の説明】

【００６３】

１０１　構造化文書群
１０２　入力文書解析手段
１０３　要素名登録手段
１０４　祖先パス名登録手段
１０５　属性名登録手段
１０６　出現情報登録手段
１０７　要素名辞書
１０８　祖先パス名辞書
１０９　属性名辞書
１１０　出現位置索引
１１１　要素出現情報格納手段
１１２　祖先パス出現情報格納手段
１１３　属性出現情報格納手段
１１４　テキスト出現情報格納手段
１１５　検索式
１１６　検索条件入力手段
１１７　検索条件解析手段
１１８　出現情報取得手段
１１９　検索結果出力手段
１２０　検索結果

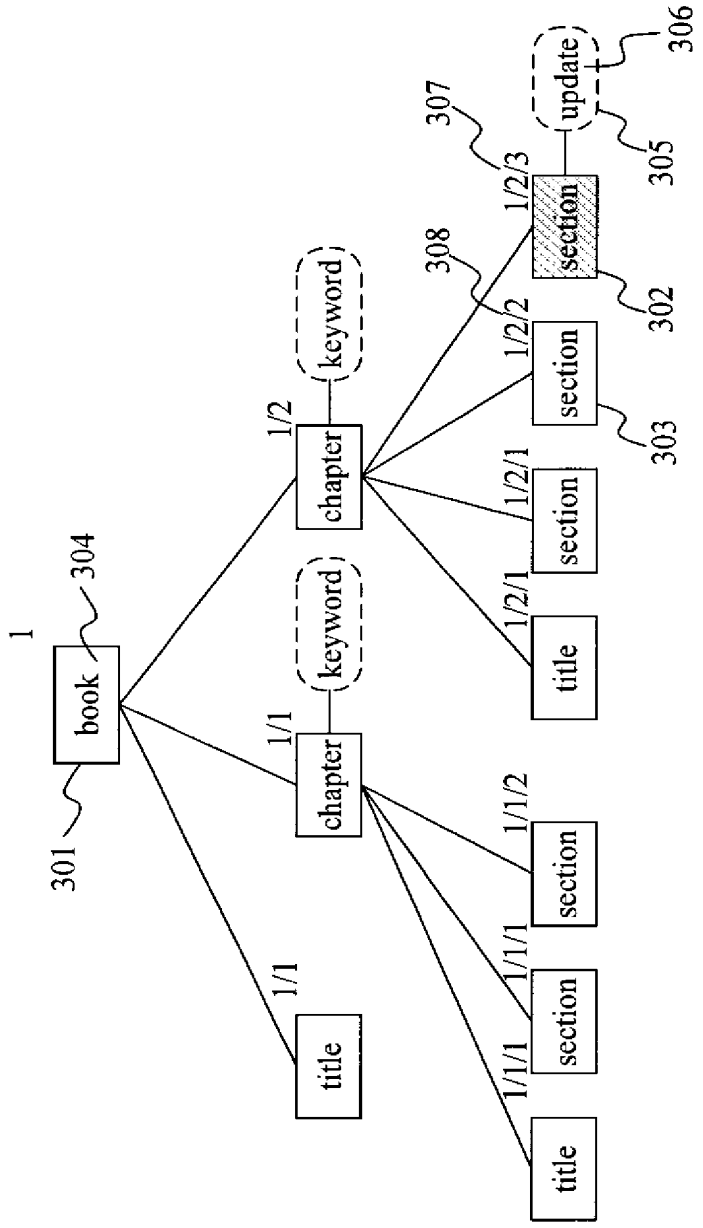
【書類名】図面
【図１】


101
構造化
文書群
102
入力文書
解析手段
要素名
登録手段
103
祖先パス名
登録手段
104
属性名
登録手段
105
出現情報
登録手段
106
107
要素名
辞書
108
祖先パス名
辞書
109
属性名
辞書
111
要素出現情報格納手段
112
祖先パス出現情報格納手段
113
属性出現情報格納手段
114
テキスト出現情報格納手段
出現位置索引
110
検索条件
解析手段
117
出現情報
取得手段
118
検索条件
入力手段
116
検索結果
出力手段
119
検索式
115
検索結果
120

【図２】

```
<book>
   <title>文書検索</title>
   <chapter keyword="歴史">
      <title>文書検索の歴史</title>
      <section>キーワード検索は、・・・</section>
      <section>その後、全文検索が・・・</section>
   </chapter>
   <chapter keyword="索引">
      <title>索引方式</title>
      <section>最長一致切り出しによる・・・</section>
      <section>ｎ－gram索引方式は・・・</section>
      <section update="200406">新たに極大単語索引方式が・・・</section>
   </chapter>
</book>
```

【図３】




【図４】

| 要素名ID | 要素名 |
|---|---|
| 1 | book |
| 2 | title |
| 3 | chapter |
| 4 | section |

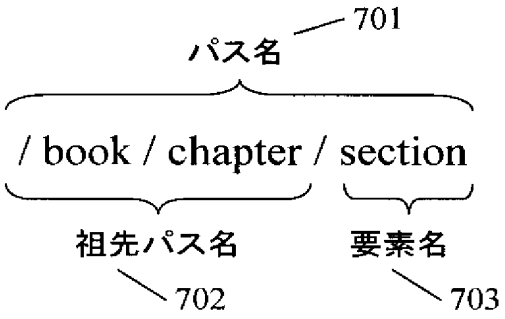
【図５】

| 祖先パス名ID | 祖先パス名 |
|---|---|
| 1 | / |
| 2 | /book |
| 3 | /book/chapter |

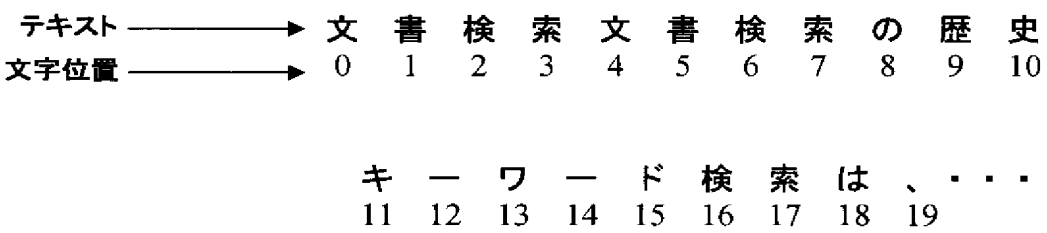
【図６】

| 属性名ID | 属性名 |
|---|---|
| 1 | keyword |
| 2 | update |

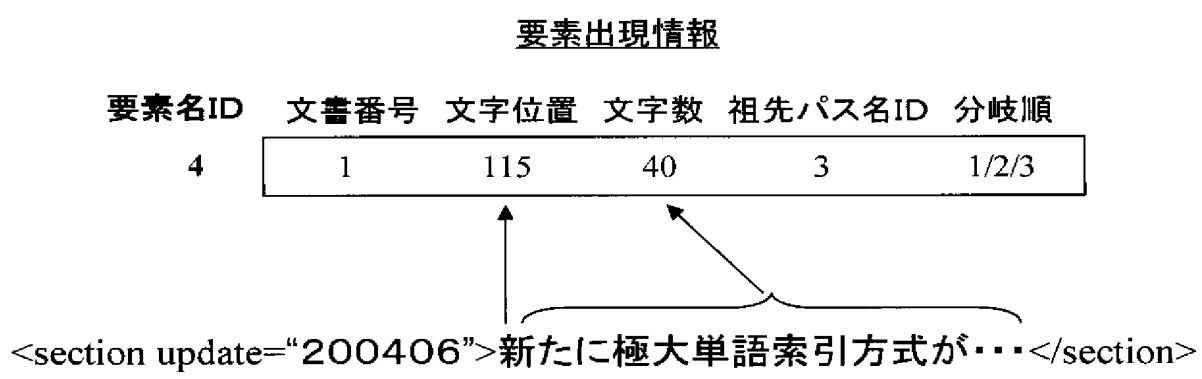
【図７】




【図８】




【図９】



【図１０】

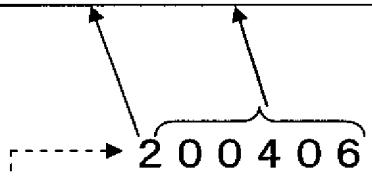
祖先パス出現情報

| 祖先パス名ID | 文書番号 | 文字位置 | 文字数 | 要素名ID | 分岐順 |
|---|---|---|---|---|---|
| 3 | 1 | 115 | 40 | 4 | 1/2/3 |

【図１１】

属性出現情報

| 属性名ID | 文書番号 | 文字位置 | 文字数 | 祖先パス名ID | 要素名ID | 分岐順 |
|---|---|---|---|---|---|---|
| 2 | 1 | 115 | 6 | 3 | 4 | 1/2/3 |




<section update="200406">新たに極大単語索引方式が・・・</section>

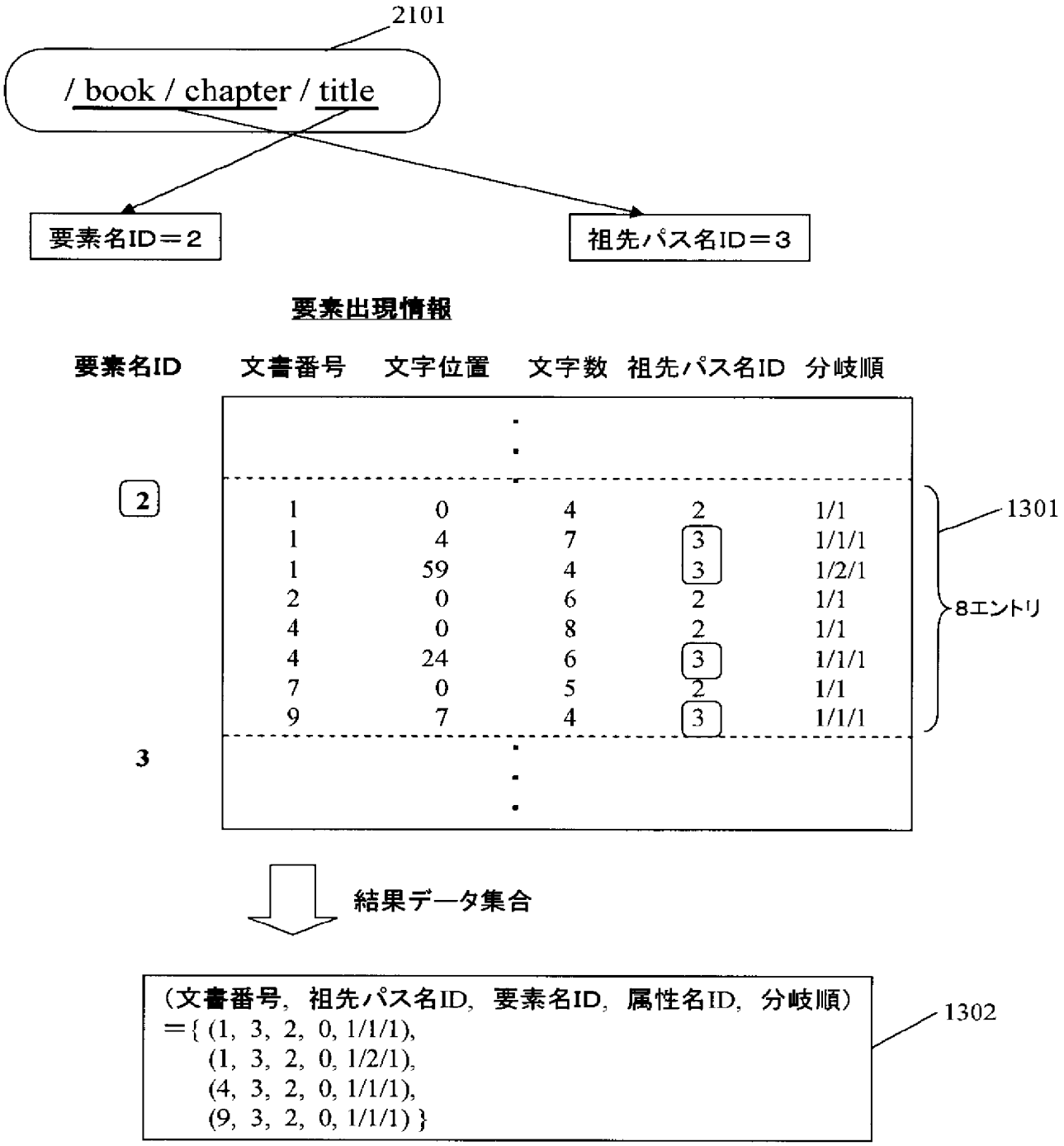
【図１２】

テキスト出現情報

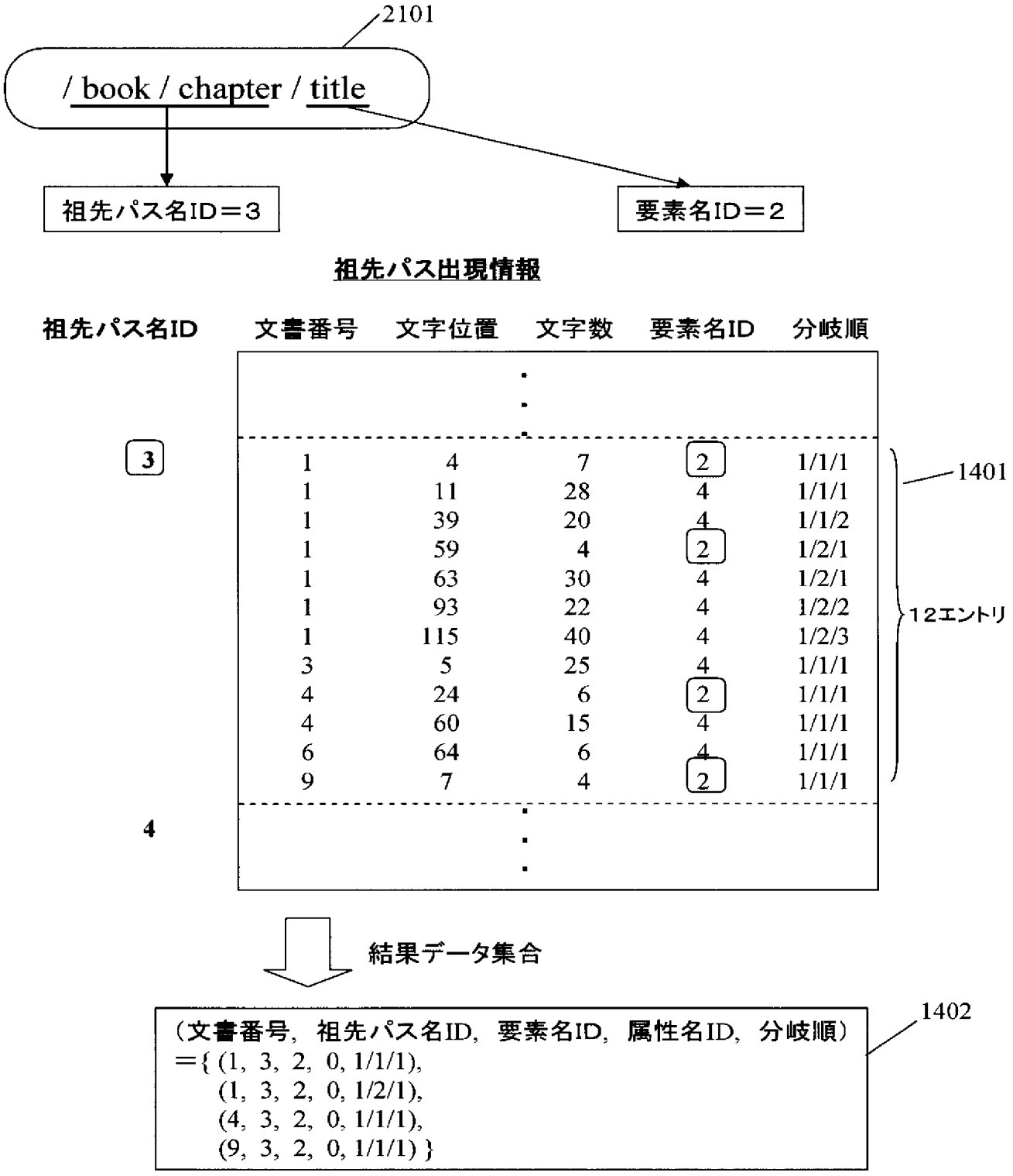
| | 部分文字列 | 文書番号 | 文字位置 | 祖先パス名ID | 要素名ID | 属性名ID | 分岐順 |
|---|---|---|---|---|---|---|---|
| | “新た” | 1 | 115 | 3 | 4 | 0 | 1/2/3 |
| | “たに” | 1 | 116 | 3 | 4 | 0 | 1/2/3 |
| | “に極” | 1 | 117 | 3 | 4 | 0 | 1/2/3 |
| 1201 | “極大” | 1 | 118 | 3 | 4 | 0 | 1/2/3 |
| | “大単” | 1 | 119 | 3 | 4 | 0 | 1/2/3 |
| | “単語” | 1 | 120 | 3 | 4 | 0 | 1/2/3 |
| | “２０” | 1 | 115 | 3 | 4 | **2** | 1/2/3 |
| 1202 | “００” | 1 | 116 | 3 | 4 | **2** | 1/2/3 |
| | “０４” | 1 | 117 | 3 | 4 | **2** | 1/2/3 |
| | “４０” | 1 | 118 | 3 | 4 | **2** | 1/2/3 |
| | “０６” | 1 | 119 | 3 | 4 | **2** | 1/2/3 |

属性値でない場合は0
属性値の場合は属性名ID(≠0)

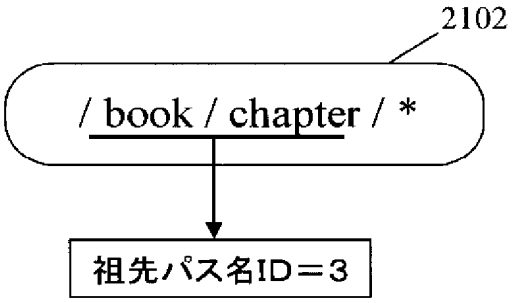
【図１３】


2101
/ book / chapter / title
要素名ID＝２
祖先パス名ID＝３
要素出現情報
要素名ID　文書番号　文字位置　文字数　祖先パス名ID　分岐順
2
1 0 4 2 1/1
1 4 7 3 1/1/1
1 59 4 3 1/2/1
2 0 6 2 1/1
4 0 8 2 1/1
4 24 6 3 1/1/1
7 0 5 2 1/1
9 7 4 3 1/1/1
3
1301
8エントリ
結果データ集合
（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{ (1, 3, 2, 0, 1/1/1),
(1, 3, 2, 0, 1/2/1),
(4, 3, 2, 0, 1/1/1),
(9, 3, 2, 0, 1/1/1) }
1302

【図１４】


2101
/ book / chapter / title
祖先パス名ID＝３
要素名ID＝２
祖先パス出現情報
祖先パス名ID
文書番号
文字位置
文字数
要素名ID
分岐順
3
1 4 7 2 1/1/1
1 11 28 4 1/1/1
1 39 20 4 1/1/2
1 59 4 2 1/2/1
1 63 30 4 1/2/1
1 93 22 4 1/2/2
1 115 40 4 1/2/3
3 5 25 4 1/1/1
4 24 6 2 1/1/1
4 60 15 4 1/1/1
6 64 6 4 1/1/1
9 7 4 2 1/1/1
4
1401
１２エントリ
結果データ集合
（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{ (1, 3, 2, 0, 1/1/1),
(1, 3, 2, 0, 1/2/1),
(4, 3, 2, 0, 1/1/1),
(9, 3, 2, 0, 1/1/1) }
1402

【図１５】




**祖先パス出現情報**

| 祖先パス名ID | 文書番号 | 文字位置 | 文字数 | 要素名ID | 分岐順 |
|---|---|---|---|---|---|
| | ： | | | | |
| 3 | 1 | 4 | 7 | 2 | 1/1/1 |
| | 1 | 11 | 28 | 4 | 1/1/1 |
| | 1 | 39 | 20 | 4 | 1/1/2 |
| | 1 | 59 | 4 | 2 | 1/2/1 |
| | 1 | 63 | 30 | 4 | 1/2/1 |
| | 1 | 93 | 22 | 4 | 1/2/2 |
| | 1 | 115 | 40 | 4 | 1/2/3 |
| | 3 | 5 | 25 | 4 | 1/1/1 |
| | 4 | 24 | 6 | 2 | 1/1/1 |
| | 4 | 60 | 15 | 4 | 1/1/1 |
| | 6 | 64 | 6 | 4 | 1/1/1 |
| | 9 | 7 | 4 | 2 | 1/1/1 |
| 4 | ： | | | | |

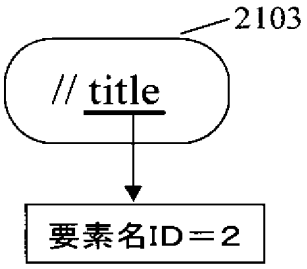
1501

結果データ集合

（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{ (1, 3, 2, 0, 1/1/1),
(1, 3, 4, 0, 1/1/1),
・・・
(6, 3, 4, 0, 1/1/1),
(9, 3, 2, 0, 1/1/1) }

1502

【図１６】




**要素出現情報**

| 要素名ID | 文書番号 | 文字位置 | 文字数 | 祖先パス名ID | 分岐順 |
|---|---|---|---|---|---|
| | ： | | | | |
| 2 | 1 | 0 | 4 | 2 | 1/1 |
| | 1 | 4 | 7 | 3 | 1/1/1 |
| | 1 | 59 | 4 | 3 | 1/2/1 |
| | 2 | 0 | 6 | 2 | 1/1 |
| | 4 | 0 | 8 | 2 | 1/1 |
| | 4 | 24 | 6 | 3 | 1/1/1 |
| | 7 | 0 | 5 | 2 | 1/1 |
| | 9 | 7 | 4 | 3 | 1/1/1 |
| 3 | ： | | | | |

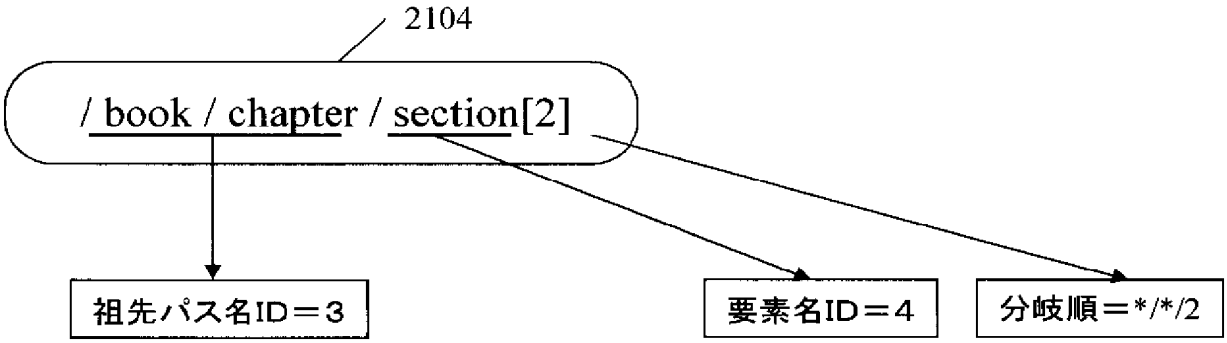
1601

結果データ集合

（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{ (1, 2, 2, 0, 1/1 ),
(1, 3, 2, 0, 1/1/1),
・・・
(7, 2, 2, 0, 1/1 ),
(9, 3, 2, 0, 1/1/1) }

1602

【図１７】




**祖先パス出現情報**

| 祖先パス名ID | 文書番号 | 文字位置 | 文字数 | 要素名ID | 分岐順 |
|---|---|---|---|---|---|
| | | ： | | | |
| 3 | 1 | 4 | 7 | 2 | 1/1/1 |
| | 1 | 11 | 28 | 4 | 1/1/1 |
| | 1 | 39 | 20 | 4 | 1/1/2 |
| | 1 | 59 | 4 | 2 | 1/2/1 |
| | 1 | 63 | 30 | 4 | 1/2/1 |
| | 1 | 93 | 22 | 4 | 1/2/2 |
| | 1 | 115 | 40 | 4 | 1/2/3 |
| | 3 | 5 | 25 | 4 | 1/1/1 |
| | 4 | 24 | 6 | 2 | 1/1/1 |
| | 4 | 60 | 15 | 4 | 1/1/1 |
| | 6 | 64 | 6 | 4 | 1/1/1 |
| | 9 | 7 | 4 | 2 | 1/1/1 |
| 4 | | ： | | | |

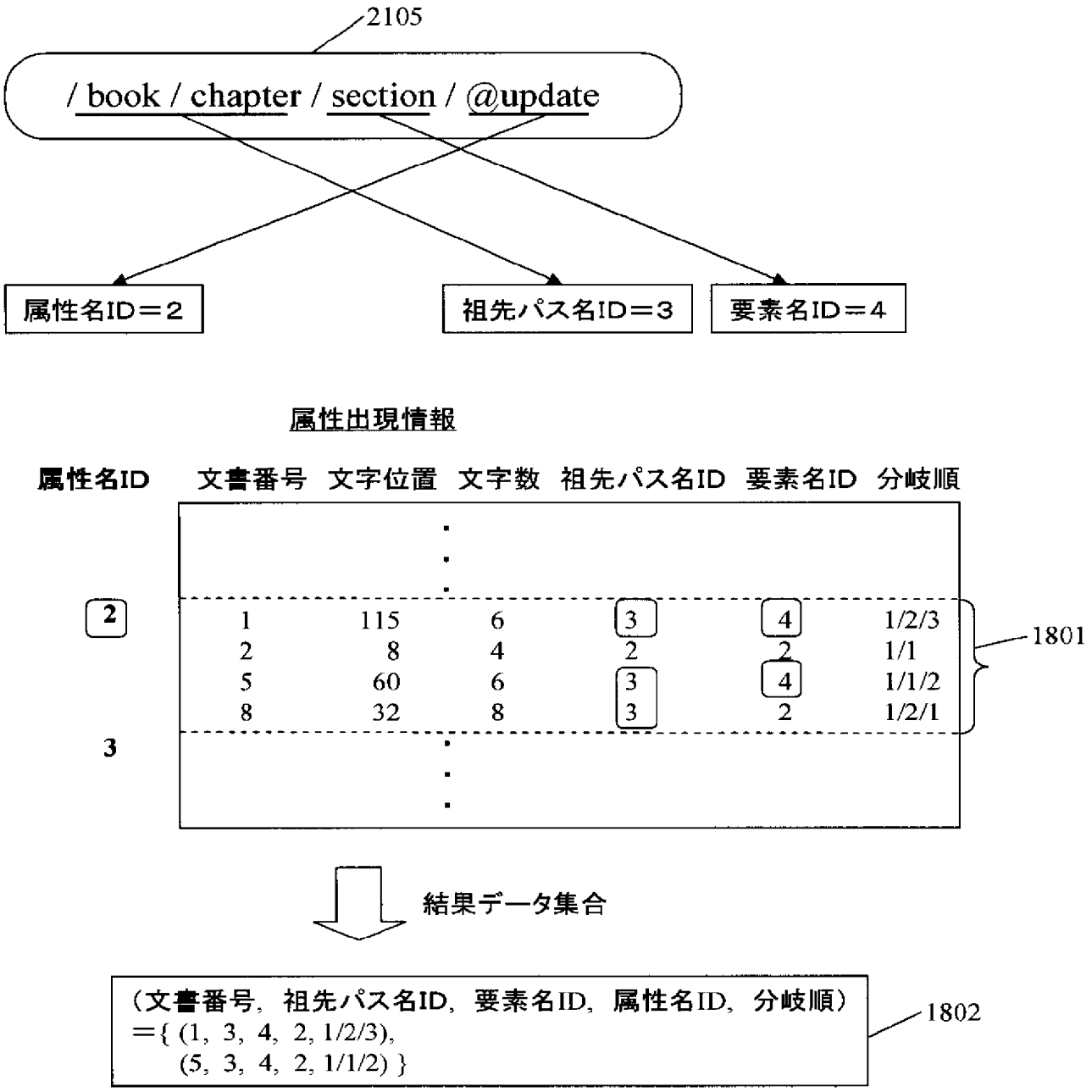
1701

結果データ集合

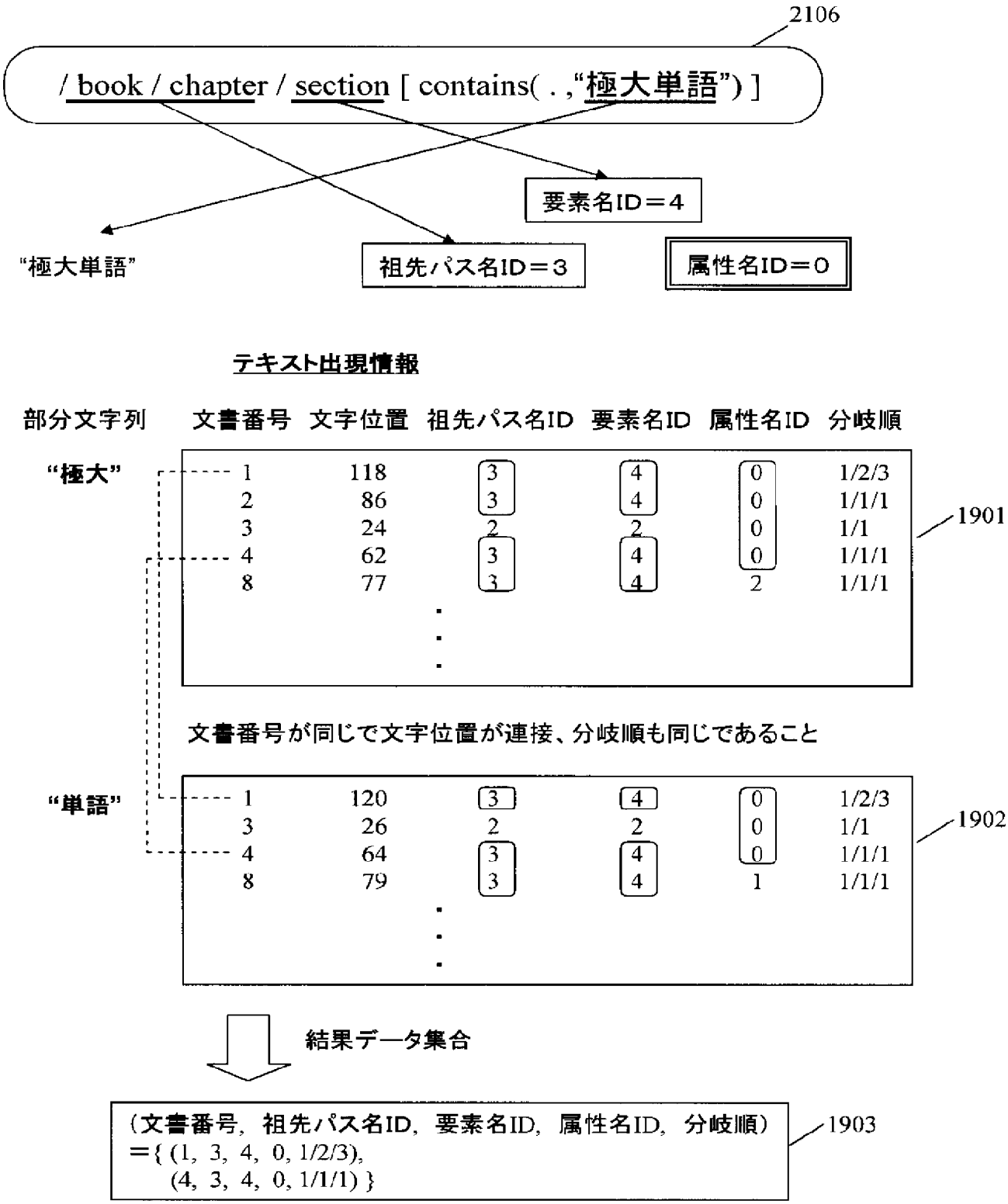
（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{ (1, 3, 4, 0, 1/1/2),
(1, 3, 4, 0, 1/2/2) }

1702

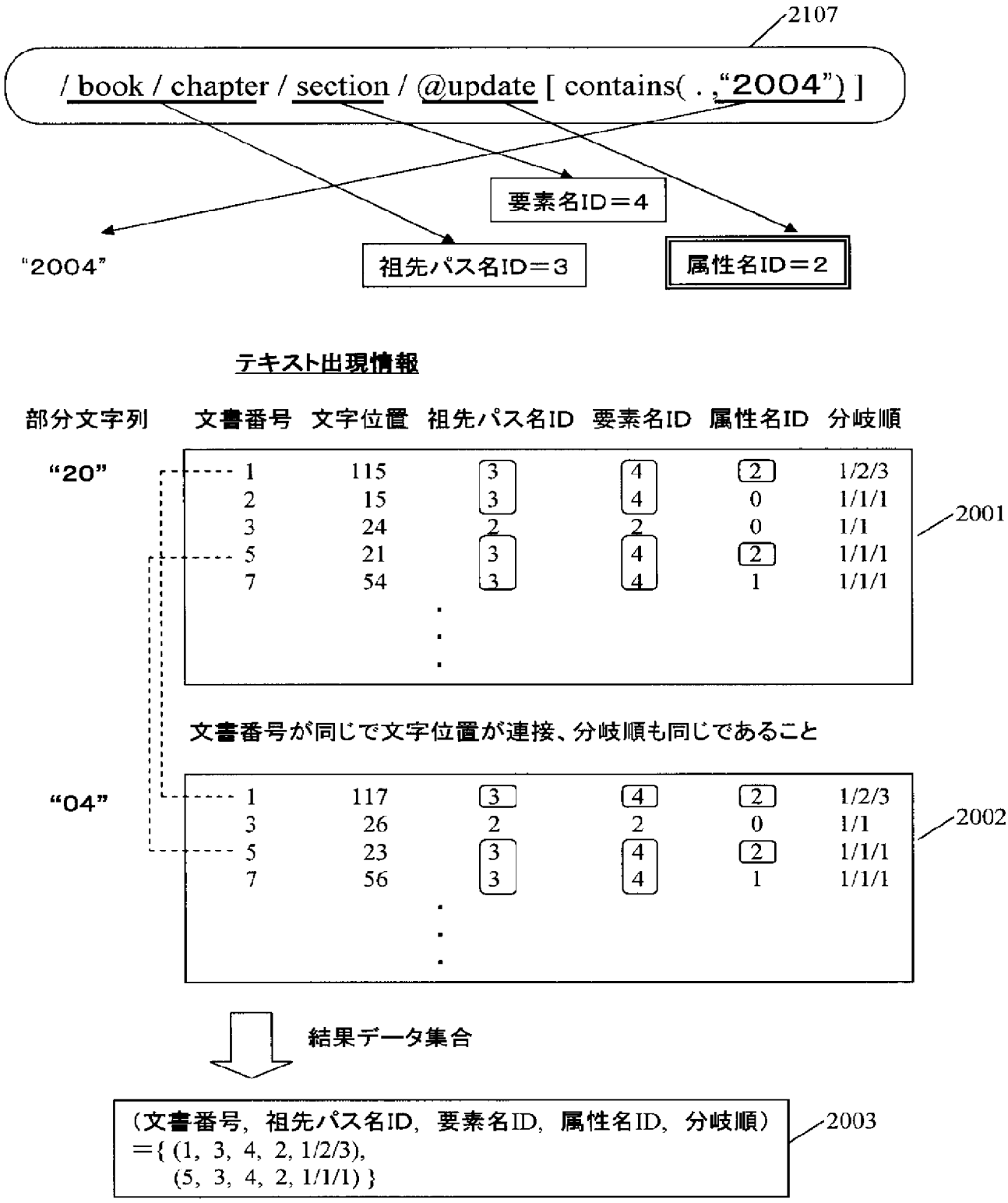
【図１８】


2105
/ book / chapter / section / @update
属性名ID＝２
祖先パス名ID＝３
要素名ID＝４
属性出現情報
属性名ID 文書番号 文字位置 文字数 祖先パス名ID 要素名ID 分岐順
2
1 115 6 3 4 1/2/3
2 8 4 2 2 1/1
5 60 6 3 4 1/1/2
8 32 8 3 2 1/2/1
3
1801
結果データ集合
（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{ (1, 3, 4, 2, 1/2/3),
(5, 3, 4, 2, 1/1/2) }
1802

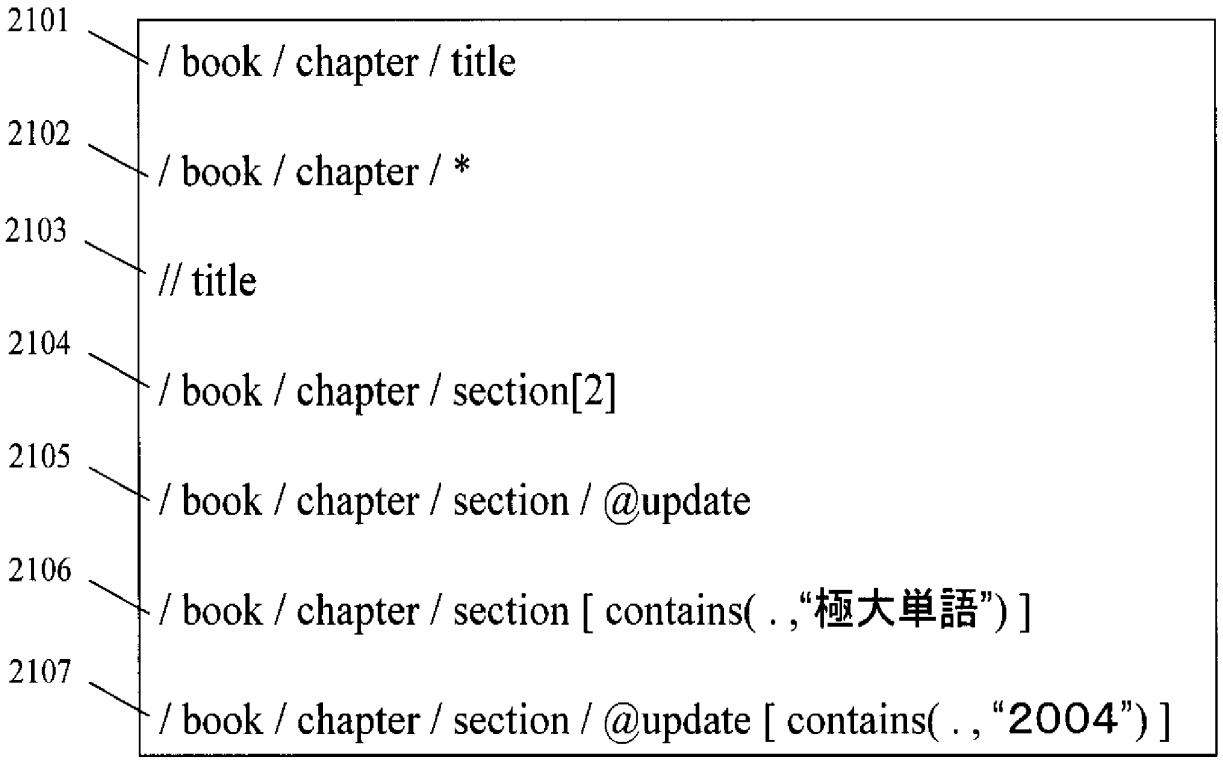
【図１９】


2106
/ book / chapter / section [ contains( . ,“極大単語”) ]
要素名ID＝4
“極大単語”
祖先パス名ID＝3
属性名ID＝0
テキスト出現情報
部分文字列 文書番号 文字位置 祖先パス名ID 要素名ID 属性名ID 分岐順
“極大”
1 118 3 4 0 1/2/3
2 86 3 4 0 1/1/1
3 24 2 2 0 1/1
4 62 3 4 0 1/1/1
8 77 3 4 2 1/1/1
1901
文書番号が同じで文字位置が連接、分岐順も同じであること
“単語”
1 120 3 4 0 1/2/3
3 26 2 2 0 1/1
4 64 3 4 0 1/1/1
8 79 3 4 1 1/1/1
1902
結果データ集合
（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
={ (1, 3, 4, 0, 1/2/3),
(4, 3, 4, 0, 1/1/1) }
1903

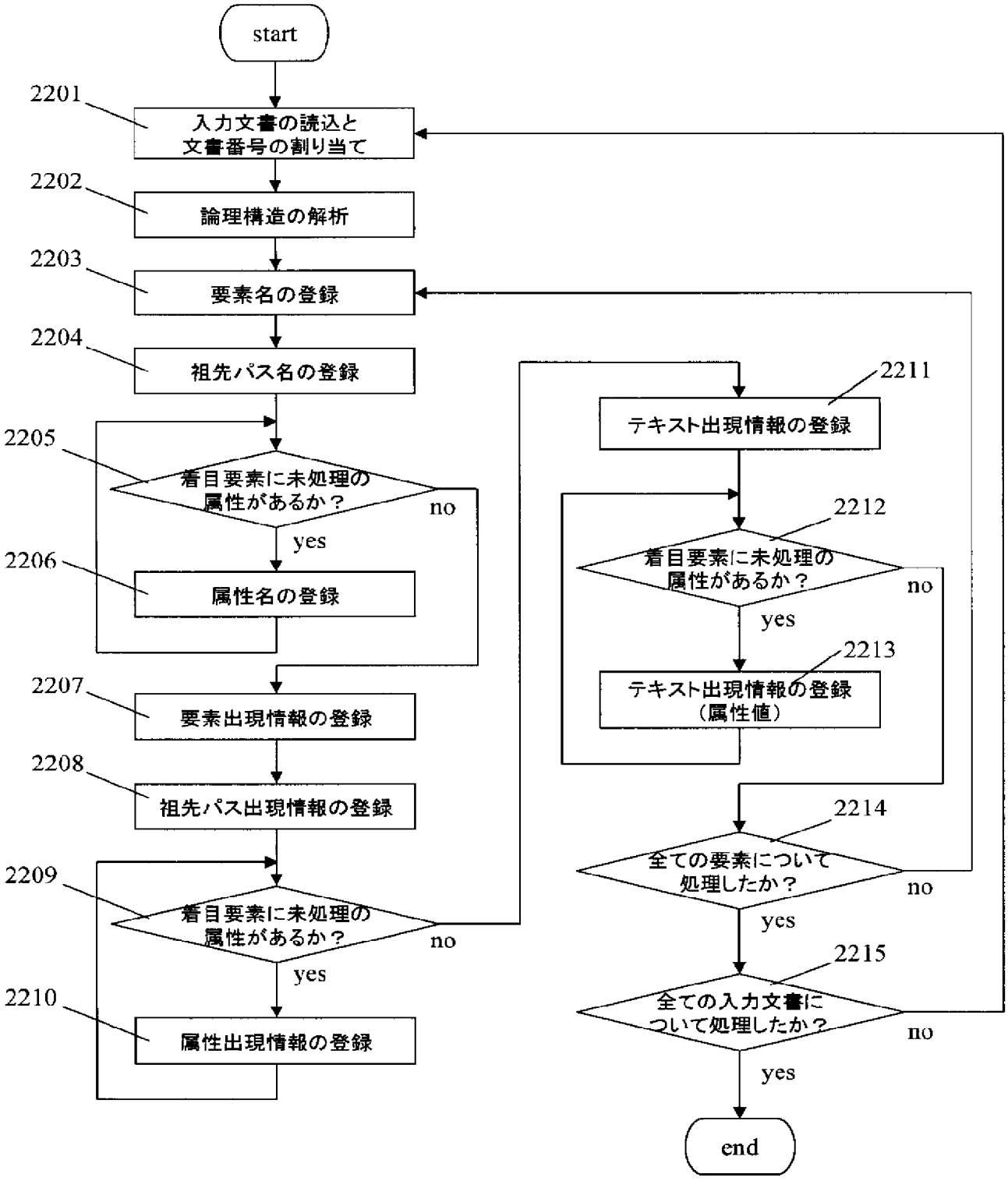
【図２０】


2107
/ book / chapter / section / @update [ contains( . ,"2004") ]
要素名ID＝4
"2004"
祖先パス名ID＝3
属性名ID＝2
テキスト出現情報
部分文字列
文書番号
文字位置
祖先パス名ID
要素名ID
属性名ID
分岐順
"20"
1 115 3 4 2 1/2/3
2 15 3 4 0 1/1/1
3 24 2 2 0 1/1
5 21 3 4 2 1/1/1
7 54 3 4 1 1/1/1
2001
文書番号が同じで文字位置が連接、分岐順も同じであること
"04"
1 117 3 4 2 1/2/3
3 26 2 2 0 1/1
5 23 3 4 2 1/1/1
7 56 3 4 1 1/1/1
2002
結果データ集合
（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
={ (1, 3, 4, 2, 1/2/3),
(5, 3, 4, 2, 1/1/1) }
2003

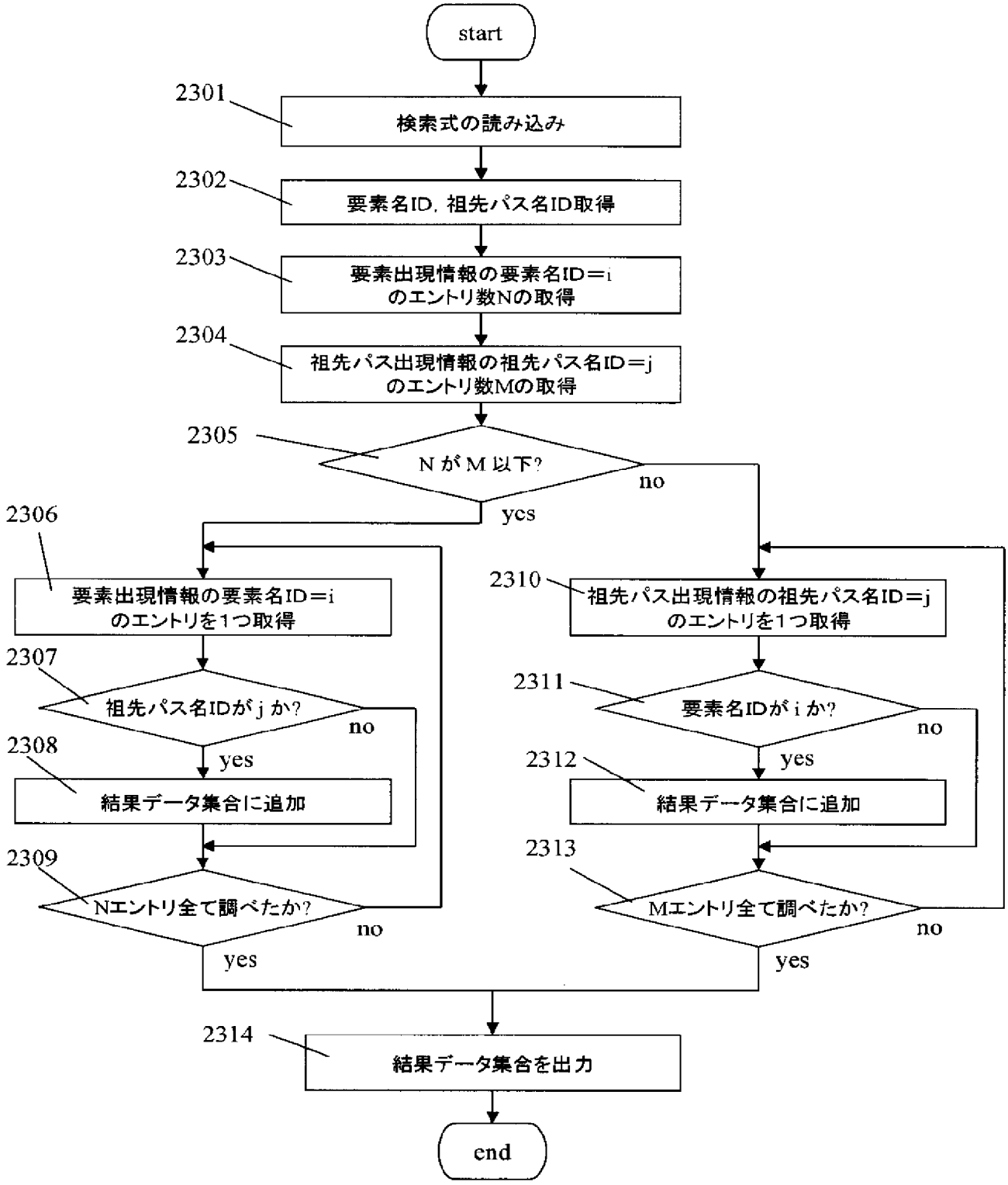
【図２１】


2101
/ book / chapter / title
2102
/ book / chapter / *
2103
// title
2104
/ book / chapter / section[2]
2105
/ book / chapter / section / @update
2106
/ book / chapter / section [ contains( . ,“極大単語”) ]
2107
/ book / chapter / section / @update [ contains( . , “2004”) ]

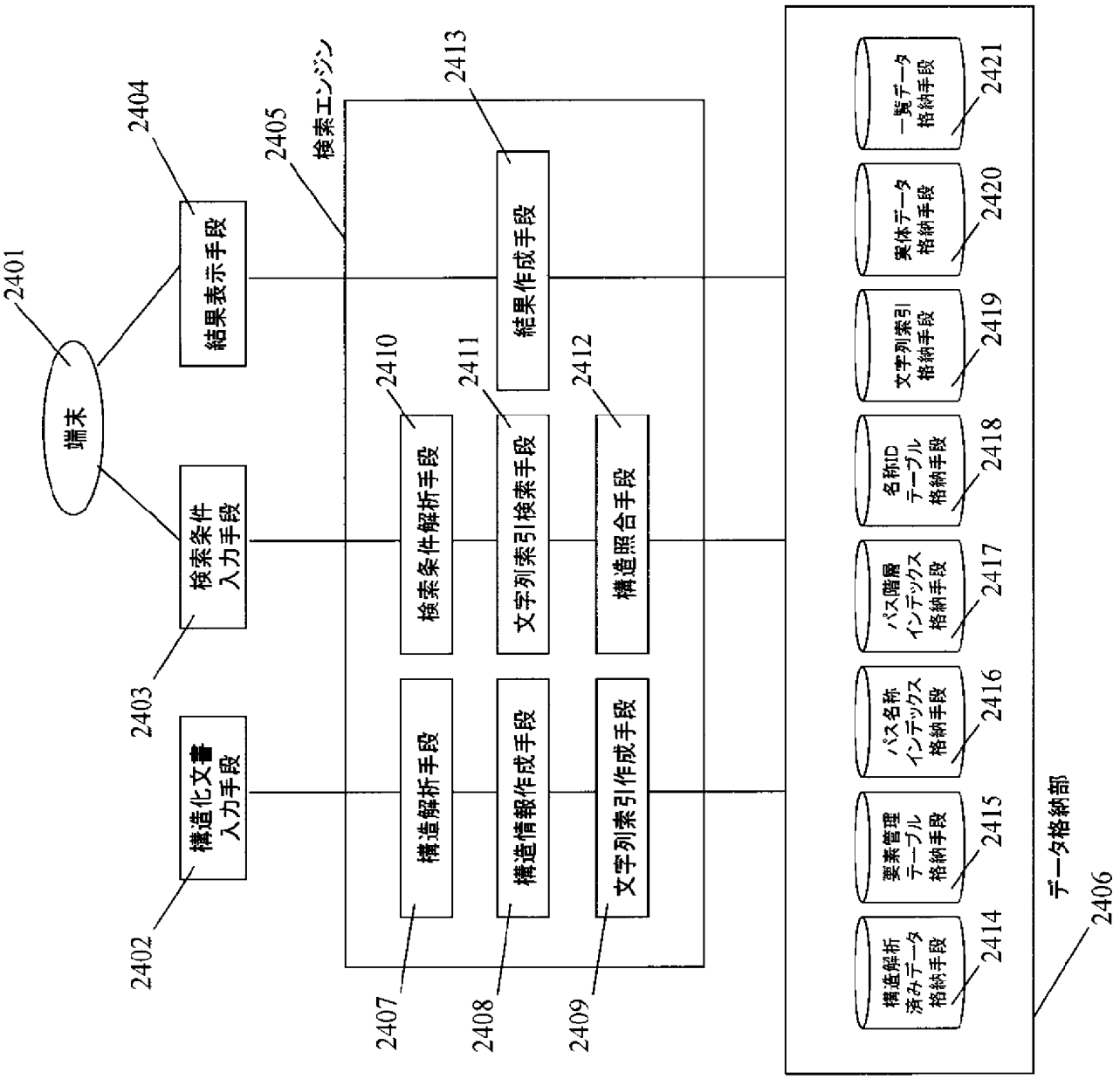
【図２２】


start
2201
入力文書の読込と
文書番号の割り当て
2202
論理構造の解析
2203
要素名の登録
2204
祖先パス名の登録
2205
着目要素に未処理の
属性があるか？
no
yes
2206
属性名の登録
2207
要素出現情報の登録
2208
祖先パス出現情報の登録
2209
着目要素に未処理の
属性があるか？
no
yes
2210
属性出現情報の登録
2211
テキスト出現情報の登録
2212
着目要素に未処理の
属性があるか？
no
yes
2213
テキスト出現情報の登録
（属性値）
2214
全ての要素について
処理したか？
no
yes
2215
全ての入力文書に
ついて処理したか？
no
yes
end

【図２３】


start
2301
検索式の読み込み
2302
要素名ID，祖先パス名ID取得
2303
要素出現情報の要素名ID＝i
のエントリ数Nの取得
2304
祖先パス出現情報の祖先パス名ID＝j
のエントリ数Mの取得
2305
NがM以下?
no
yes
2306
要素出現情報の要素名ID＝i
のエントリを1つ取得
2307
祖先パス名IDがjか?
no
yes
2308
結果データ集合に追加
2309
Nエントリ全て調べたか?
no
yes
2310
祖先パス出現情報の祖先パス名ID＝j
のエントリを1つ取得
2311
要素名IDがiか?
no
yes
2312
結果データ集合に追加
2313
Mエントリ全て調べたか?
no
yes
2314
結果データ集合を出力
end

【図２４】


2401
端末
2402
構造化文書入力手段
2403
検索条件入力手段
2404
結果表示手段
2405
検索エンジン
2407
構造解析手段
2408
構造情報作成手段
2409
文字列索引作成手段
2410
検索条件解析手段
2411
文字列索引検索手段
2412
構造照合手段
2413
結果作成手段
2414
構造解析済みデータ格納手段
2415
要素管理テーブル格納手段
2416
パス名称インデックス格納手段
2417
パス階層インデックス格納手段
2418
名称IDテーブル格納手段
2419
文字列索引格納手段
2420
実体データ格納手段
2421
一覧データ格納手段
2406
データ格納部

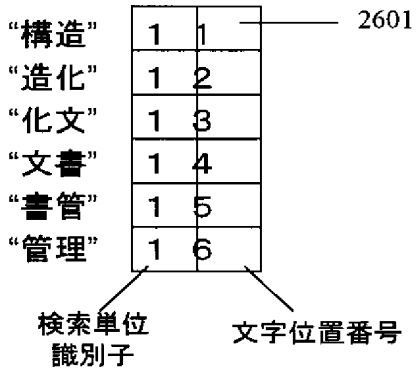
【図２５】

| 検索単位識別子 | 文書番号 | パス名称ID | パス階層ID | 名称ID |
|---|---|---|---|---|
| 1 | 1 | N2 | L2 | T3 |
| 2 | 1 | N3 | L2 | T4 |
| 3 | 1 | N3 | L6 | T4 |
| 4 | 1 | N4 | L2 | T5 |
| 5 | 1 | N7 | L3 | T8 |
| 6 | 1 | N8 | L3 | T9 |
| 7 | 1 | N10 | L4 | T11 |
| 8 | 1 | N11 | L4 | T9 |
| ・<br>・<br>・ | … | … | … | … |

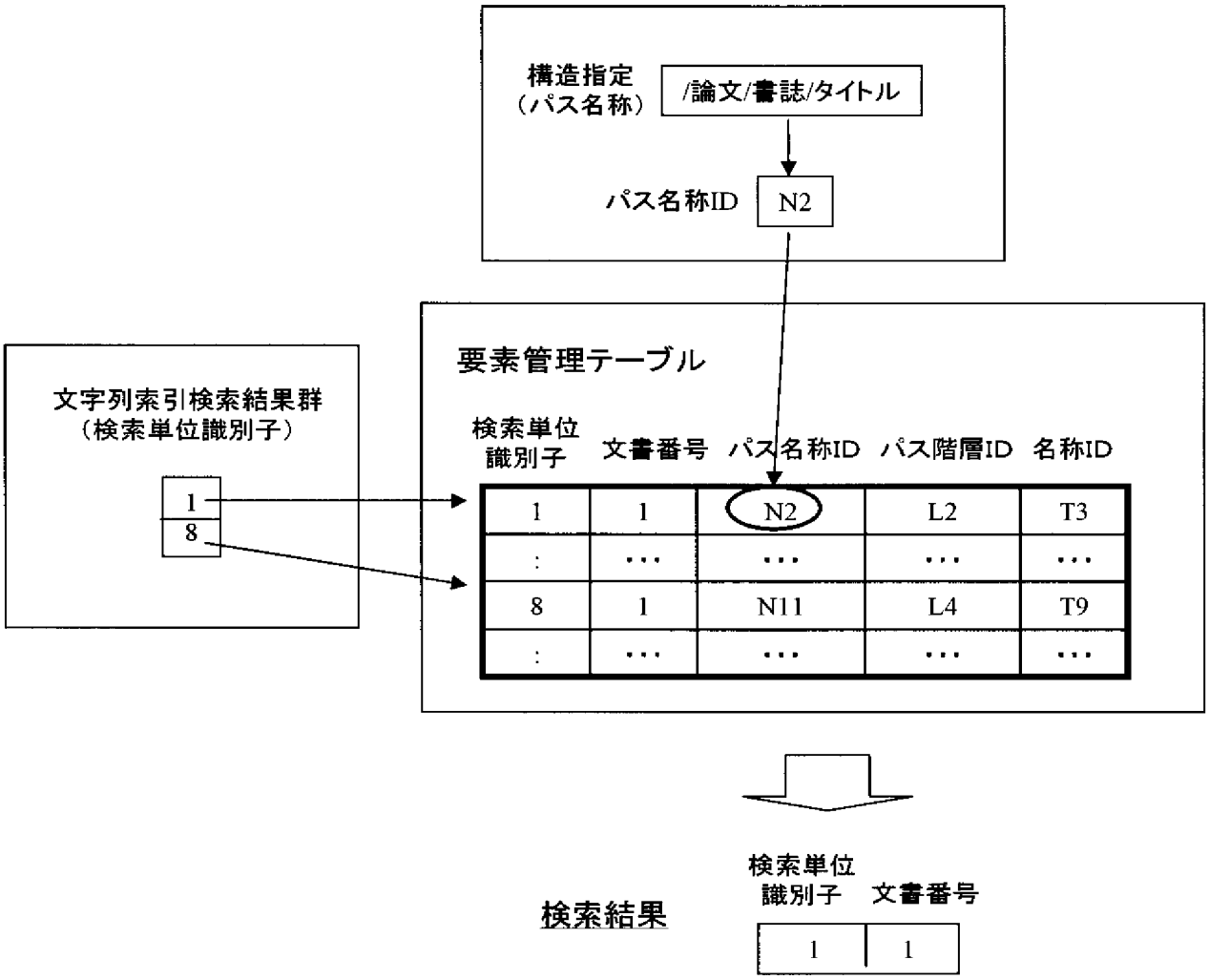
要素管理テーブル

【図２６】

<タイトル>構造化文書管理</タイトル>



【図２７】


構造指定
（パス名称）
/論文/書誌/タイトル
パス名称ID
N2
要素管理テーブル
文字列索引検索結果群
（検索単位識別子）
1
8
検索単位
識別子
文書番号
パス名称ID
パス階層ID
名称ID
1
1
N2
L2
T3
:
・・・
・・・
・・・
・・・
8
1
N11
L4
T9
:
・・・
・・・
・・・
・・・
検索結果
検索単位
識別子
文書番号
1
1

【書類名】要約書
【要約】
【課題】様々な検索条件での構造化文書に対する検索を効率よく行う。構造条件のみでの検索や、属性値に対する文字列検索ができるようにすることを目的とする。
【解決手段】要素の出現情報を、要素名ＩＤをキーにして格納した要素出現情報格納手段と、要素の出現情報を、その要素の祖先パス名ＩＤをキーにして格納した祖先パス出現情報格納手段と、属性の出現情報を、属性名ＩＤをキーにして格納した属性出現情報格納手段と、要素実体のテキスト文字列、および要素のもつ属性の属性値に関する出現情報を、部分文字列をキーにして格納したテキスト出現情報格納手段とを備える。
【選択図】図１

出願人履歴

０００００５８２１
19900828
新規登録

大阪府門真市大字門真１００６番地
松下電器産業株式会社